COMP760, SUMMARY OF LECTURE 12.

HAMED HATAMI

1. Introduction to information theory

In 1948, Shannon in his seminal paper [Sha48] introduced information theory as a tool for studying the communication cost of transmission tasks. In this setting, Alice receives a random input $X \sim (\Omega, p)$ and wants to transmit it to Bob. Shannon's source coding theorem (Theorem 2 below) concerns the case where Alice is sending many copies of X. More precisely, she receives $(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are i.i.d. copies of X, and she wants to send a coded version of this $f(X_1, \ldots, X_n) \in \{0,1\}^{\mathbb{N}}$ to Bob, so that Bob, with probability at least $1-\epsilon$, can successfully decode it using his function g:

$$\Pr[g(f(X_1, \ldots, X_n))) \neq (X_1, \ldots, X_n)] \leq \epsilon,$$

where $\epsilon > 0$ is a fixed small number. Here Alice wants to optimize the number of bits that she is sending $\max |f(x_1, \ldots, x_n)|$, where $|f(x_1, \ldots, x_n)|$ denotes the length of the string. Let $C_n(X, \epsilon)$ denote this number. Note that

$$\Pr[(X_1, \ldots, X_n) = (a_1, \ldots, a_n)] = p(a_1) \ldots p(a_n) = 2^{\log p(a_1) + \ldots + \log p(a_n)}.$$

Let us investigate the power. What is its expected value when $a_1, \ldots, a_n$ are the values of $X_1, \ldots, X_n$?

$$\mathbb{E}[\log p(a_1) + \ldots + \log p(a_n)] = n\mathbb{E}[\log p(a_1)] = -n \sum_{\substack{a \in \Omega \\ p(a) \neq 0}} p(a) \log \frac{1}{p(a)} = -nH(X),$$

where $H(X)$ is called the entropy of X.

Definition 1 (Entropy). *The entropy of a random variable $X \sim (\Omega, p)$ is defined as*

$$H(X) := \sum_{\substack{a \in \Omega \\ p(a) \neq 0}} p(a) \log \frac{1}{p(X(a))} = \mathbb{E} \log \frac{1}{p(X)}.$$

It follows from Law of Large Numbers that for a random $(X_1, \ldots, X_n)$ the value of

$$\log \Pr[(X_1, \ldots, X_n) = (a_1, ..., a_n)]$$

is highly concentrated around its expected value $-nH(X)$. This shows that if we define

$$A_n^\epsilon = \{(a_1, \ldots, a_n) \ : \ |\log \Pr[(a_1, \ldots, a_n)] + nH(X)| \leq \epsilon n\},$$

then A_n^ϵ is almost of full measure, or in other words

$$\Pr[(X_1, \ldots, X_n) \in A_n^\epsilon] \geq 1 - \epsilon.$$

Note that A_n^ϵ is precisely the set of all points $(a_1, \ldots, a_n) \in \Omega^n$ with

$$2^{-nH(X) - \epsilon n} \leq \Pr[(a_1, \ldots, a_n)] \leq 2^{-nH(X) + \epsilon n}.$$

Here intuitively you can think that these points are almost uniformly distributed on A_n^ϵ (ignoring the multiplicative factors $2^{O(\pm\epsilon n)}$). Hence we roughly know the size of A_n^ϵ,

$$2^{nH(X)-\epsilon n-1} \leq (1-\epsilon)2^{nH(X)-\epsilon n} \leq |A_n^\epsilon| \leq 2^{nH(X)+\epsilon n}.$$

It follows from the upper-bound that one can assign a unique binary string $f(a_1,\ldots,a_n)$ of length at most $nH(X)+\epsilon n$ to each point $(a_1,\ldots,a_n) \in A_n^\epsilon$. Hence the following protocol satisfies our requirement and shows that $C_n(X,\epsilon) \leq nH(X)+\epsilon n$.

- Alice checks to see if $(X_1,\ldots,X_n) \in A_n^\epsilon$:
  - If YES she sends $f(X_1,\ldots,X_n)$ to Bob
  - If NO she sends "FAILED" to Bob.

On the other hand note that since $2^{nH(X)-\epsilon n-1} \leq |A_n^\epsilon|$ any protocol that uses less than say $nH(X)-2\epsilon n$ bits can get at most $2^{nH(X)-2\epsilon n}$ points of A_n^ϵ right, and that has probability mass at most

$$2^{nH(X)-2\epsilon n} \times 2^{-nH(X)+\epsilon n} \leq 2^{\epsilon n} \leq \epsilon,$$

where we used the bound $\Pr[(a_1,\ldots,a_n)] \leq 2^{-nH(X)+\epsilon n}$ on the points $(a_1,\ldots,a_n) \in A_n^\epsilon$. So such a protocol at the best case can get the points out of A_n^ϵ plus this $\geq \epsilon$ measure points in A_n^ϵ correct. Hence the probability of error is going to be at least $1-2\epsilon > \epsilon$, and this shows $C_n(X,\epsilon) \geq nH(X)-2\epsilon n$. Summarizing this we proved

$$nH(X)-2\epsilon n \leq C_n(X,\epsilon) \leq nH(X)+\epsilon n,$$

which shows

$$\lim_{\epsilon\to 0}\lim_{n\to\infty} \frac{C_n(X,\epsilon)}{n} = H(X).$$

This is known as the Shannon's source coding theorem which intuitively says that n copies of X contains roughly $nH(X)$ bits of information.

Theorem 2 (Shannon's source coding theorem)**.** *The n i.i.d. copies $X_1,\ldots,X_n$ of X can be compressed into $nH(X)+o(n)$ bits with information loss $o(1)$, and conversely every compressing that uses $nH(X)-\Omega(n)$ bits has significant information loss $1-o(1)$.*

Another important theorem that shows that $H(X)$ is the "right" measure of information is Huffman's coding theorem. Assume now that Alice wants to send just one copy of X, and she wants to minimize the expected number of sent bits: $\mathbb{E}|f(X)|$. Let us denote the optimal value of this quantity with $C(X)$. This time no information loss is allowed. Thus $g(f(x)) = x$ for all $x \in \Omega$ with $p(x) > 0$. You can also think that she is sending many copies of X, one at a time, and in long run wants to optimize the number of bits that she is sending (this is going to be concentrated around $n\mathbb{E}[|f(X)|]$ after n rounds.)

Huffman's coding theorem says that for optimal f we have

$$H(X) \leq \mathbb{E}[|f(X)|] \leq H(X)+1.$$

To prove the upper-bound one can construct a tree by starting first from leaves, labeled with points $a \in \Omega$ with weight $p(a) > 0$. All these nodes are active in the begging. Then at every step one takes two active nodes with the smallest weights and create an active parent for them with the weight that is sum of the weights of the two children. The two children become deactivated. We continue this process until we end up with one node which will be the root of a binary tree. The paths from the root to leaves will correspond to $0,1$-strings (say 0 every time one goes to a left child and 1 every time one goes to a right child), and it is not difficult to see that if we sample a leaf according to the probability $p(\cdot)$, then the expected height of that leaf is at most $H(X)+1$. We will not give a detailed proof of this theorem as it is not directly related to the topic of the course. However, the interested reader can refer to [CT06] for a proof. Now not that if we define $C_n(X) := C(X_1,\ldots,X_n)$ where $X_1,\ldots,X_n$ are independent copies of X, then by Huffman's coding theorem, we have

$$H(X_1,\ldots,X_n) \le C_n(X) \le H(X_1,\ldots,X_n)+1.$$

As we shall see below, since $X_1,\ldots,X_n$ are indepdendent $H(X_1,\ldots,X_n) = nH(X)$, and thus

$$\lim \frac{C_n(X)}{n} = H(X).$$

Thus Shannon's source coding theorem and Huffman's coding theorem show that $H(X)$ is the right notion for capturing the amount of information that a random variable X contains.

Let us mention some notation and some basic properties of the entropy function.

- $H(X) \ge 0$ always and $H(X) = 0$ iff $\Pr[X = a] = 1$ for some $x \in \Omega$.
- For $0 \le \alpha \le 1$ let $H(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha)\log\frac{1}{1-\alpha} \in [0,1]$ denote the entropy of the Bernoulli random variable B with $\Pr[B=1] = \alpha$ and $\Pr[B=0] = 1-\alpha$.
- Note that $H(\alpha) = 1$ if and only if $\alpha = \frac{1}{2}$.
- $H(\alpha) = H(1-\alpha)$ and $H(0) = H(1) = 0$.
- Using Taylor expansion for $\log(1-\epsilon)$, for small $\epsilon > 0$ we have
$$\epsilon \log\frac{1}{\epsilon} \le H(\epsilon) \le (1+o(1))\epsilon\log\frac{1}{\epsilon}.$$
- if X is uniformly supported on a set S, then
$$H(X) = \sum_{a\in S}\frac{1}{|S|}\log|S| = \log|S|.$$

Example 3. Suppose X, Y, Z are uniform random bits conditioned on $X \oplus Y \oplus Z = 0$. Note that they are pairwise independent but obviously not mutually independent as the value of Z is determined by the values of X and Y. Then XYZ (as it is the convention in information theory, this denotes the vector (X,Y,Z) and *not* the product of X,Y and Z), is uniformly distributed on 4 points, and hence

$$H(XYZ) = \log 4 = 2.$$

This demonstrates that XYZ contains only two bits of information as the value of Z is already determined by the value of X and Y. However note that $H(Z) = 1$ as Z by itself contains some information. Also the identity

$$H(XYZ) = H(XY) = 2$$

shows that XYZ does not contain any extra information than XY. ■

1.1. **Conditional Entropy.** Let's go back to Shannon's source coding theorem where Alice wants to transmit $(X_1,\ldots,X_n)$ to Bob, but now suppose that Bob has a random variable $(Y_1,\ldots,Y_n)$ where $Y_1,\ldots,Y_n$ are i.i.d copies of a random variable Y. Alice sees $Y_1,\ldots,Y_n$ and wants to use this to save in her transmission cost. Obviously if Y and X are independent then this will not be of any use, but if X and Y are correlated, then this already provides some information about $X_1,\ldots,X_n$ to Bob and thus Alice might get away with sending less information. Note that in the extreme case that Y and X are completely correlated $Y := X$, then Alice does not need to send any information to Bob as Bob already knows $(X_1, \ldots, X_n) = (Y_1, \ldots, Y_n)$. Using our notation from the source coding theorem let us denote by $C_n(X|Y)$ the number of bits that Alice needs to send under these conditions to have the probability of information loss $\geq \epsilon$.

We showed that in the case $Y = X$ we have $C_n(X|Y) = 0$. Let us look at another example. Suppose $X = (B, B')$ where B and B' are two independent random bits, and $Y = B \oplus B'$. Then obviously to send $(X_1, \ldots, X_n) = (B_1B'_1, \ldots, B_nB'_n)$ to Bob, Alice only needs to send $(B_1, \ldots, B_n)$, as Bob can use this together with the information $Y = (B_1 \oplus B'_1, \ldots, B_n \oplus B'_n)$ to recover all of $(X_1, \ldots, X_n)$. Hence $C_n(X|Y) = n$ in this case, while $C_n(X) = nH(X) = 2n$. It turns out that

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} C_n(X|Y, \epsilon)$$

converges to the so called conditional entropy.

Definition 4 (Conditional Entropy)**.** *Let X and Y be two random variables with a joint distribution $p(x, y)$, then*

$$H(X|Y) = \mathbb{E}_{b \sim p}\left[H(X|Y = b)\right] = H(XY) - H(Y).$$

Note that $H(X|Y = b)$ is just the usual entropy of a random variable (the random variable that is obtained from X by conditioning on the event $Y = b$).

The condition entropy $H(X|Y)$ captures the amount of information of X left once you know Y, and the identity $H(X|Y) = H(XY) - H(Y)$ says that it is equal to the amount of information that is in X and Y together minus the amount of information that Y contains.

The following chain rule is immediate from the definition of the conditional entropy:

Theorem 5 (Chain Rule)**.** *We have*

$$H(XY) = H(X) + H(Y|X),$$

and more generally

$$H(X_1 \ldots X_n) = H(X_1) + H(X_1|X_2) + H(X_2|X_1X_2) + \ldots + H(X_n|X_1 \ldots X_{n-1}).$$

Let us recall Jansen's inequality from basic real analysis.

Theorem 6 (Jansen's inequality)**.** *If $f : C \to \mathbb{R}$ is a convex real valued function over a convex set $C \subset \mathbb{R}^d$ for some d, and X is a random variable that takes values in C, then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}\left[f(X)\right],$$

and equality holds if and only if X is constant almost everywhere.

We will use this theorem to prove the sub-additivity of entropy. Before doing so let us make a conventional notation to be used in the subsequent part of this course: Let X and Y be random variables with joint distribution $p(x, y)$. Then with a *serious abuse of notation* we shall denote the marginal distribution of X and Y respectively by $p(x)$ and $p(y)$. That is if (X, Y) is sampled according to $p(x, y)$ then

$$p(x) = \Pr[X = x] \qquad \text{and} \qquad p(y) = \Pr[Y = y].$$

Obviously a more precise notation would be to use $p_1(x)$ and $p_2(y)$ to denote these two different functions, but to avoid writing the subscripts we will use this notation.

Corollary 7 (Subadditivity of Entropy)**.** *We have*

$$H(XY) \leq H(X) + H(Y),$$

and

$$H(X|Y) \leq H(X),$$

with equalities if and only if X and Y are independent.

Proof. The two statements are equivalent as $H(X|Y) = H(XY) - H(Y)$. To prove the first statement we have

$$\begin{aligned} H(XY) - H(X) - H(Y) &= \sum_{x,y} p(x,y)\frac{1}{\log p(x,y)} - \sum_{x} p(x)\frac{1}{\log p(x)} - \sum_{y} p(y)\frac{1}{\log p(y)} \\ &= \sum_{x,y} p(x,y)\left[\frac{1}{\log p(x,y)} - \frac{1}{\log p(x)} - \frac{1}{\log p(y)}\right] \\ &= \sum_{x,y} p(x,y)\log\left(\frac{p(x)p(y)}{p(x,y)}\right) \\ &\leq \log\left(\sum_{x,y} p(x,y)\frac{p(x)p(y)}{p(x,y)}\right) = \log 1 = 0, \end{aligned}$$

where the inequality above is an application of Jansen's inequality to the function $f(z) = z\log z$. By the equality case of Theorem 6, equality holds if and only if $\frac{p(x)p(y)}{p(x,y)}$ is a constant, and since $\mathbb{E}\frac{p(x)p(y)}{p(x,y)} = 1$, we have that $p(x,y) = p(x)p(y)$ for every x and y. □

Remark 8. [Warning] Note that it is *not* true that always $H(X|Y = b) \leq H(X)$. For example let Y be a Bernoulli random variable with $\Pr[Y = 1] = \epsilon$ where ϵ is very small. Then if $Y = 1$ we sample X uniformly at random from $\{0,1\}$ and if $Y = 0$, then we deterministically set $X = 0$. In this case

$$H(X) = H(\epsilon/2) \approx \frac{\epsilon}{2}\log(2/\epsilon),$$

while

$$H(X|Y = 0) = H(1/2) = 1 > H(X).$$

Note that however, as it is expected

$$H(X|Y) = \epsilon H(1/2) + (1-\epsilon)H(0) = \epsilon < H(X).$$

■

1.2. **Some examples.** We will finish this lecture by some examples:

- Let g be a function, then $H(g(X)) \leq H(X)$ and $H(Y|X) \leq H(Y|g(X))$.

  proof: By subadditivity we have

  $$H(g(X)) \leq H(Xg(X)) = H(X),$$

  where equality is because X and $Xg(X)$ have exactly the same distribution. Similarly, as conditioning decreases the entropy:

  $$H(Y|g(X)) \geq H(Y|g(X)X) = H(Y|X).$$

■

- Consider a bin filled with n balls of various colors. In one experiment for $k \leq n$ times we take a random ball out of the bin, record its color and put it back in the bin. In the second experiment, we do not put the balls back in the bin. If $(X_1, \ldots, X_k)$ and $(Y_1, \ldots, Y_k)$ are random variables for the recorded colors in the two experiments. How do the entropy of these two random variables compare?

  Solution: Matching what intuition suggests, the first one has higher entropy. In the first experiment the colors are independent and identically distributed, and hence
  $$H(X_1, \ldots, X_k) = H(X_1) + \ldots + H(X_n) = nH(X_1)$$
  For the second experiment, $Y_1, \ldots, Y_n$ are dependent, but yet each one individually has the same distribution as X_1. Thus by subadditivity (Corollary 7)
  $$H(Y_1, \ldots, Y_k) < H(Y_1) + \ldots + H(Y_n) = nH(X_1)$$

- Let X and Y be random variables distributed according to distributions μ and ν. Let Z be a third random variable distributed according to $\lambda\mu + (1-\lambda)\nu$ where $\lambda \in [0,1]$. How does the entropy of Z compare to the entropies of X and Y?

  Solution: How can we sample Z? Let B be a Bernoulli variable with parameter λ. If $B = 1$ we set $Z := X$ and if $B = 0$, then we set $Z := Y$. Note
  $$H(Z) \geq H(Z|B) = \lambda H(Z|B=1) + (1-\lambda)H(Z|B=0) = \lambda H(X) + (1-\lambda)H(Y).$$
  On the other hand
  $$\begin{aligned} H(Z) &\leq H(BZ) = H(B) + H(Z|B) \\ &= H(\lambda) + \lambda H(X) + (1-\lambda)H(Y) \leq 1 + \lambda H(X) + (1-\lambda)H(Y). \end{aligned}$$
  So
  $$\lambda H(X) + (1-\lambda)H(Y) \leq H(Z) \leq \lambda H(X) + (1-\lambda)H(Y) + H(\lambda).$$

- Alice receives $t_1, \ldots, t_n \in \{1, 2, \ldots, 5\}^n$ which are selected uniformly and independently, and Bob has $s_1, \ldots, s_n \in \{1, 2, \ldots, 5\}^n$, visible to Alice, which are also uniform and independent but conditioned on $s_i \neq t_i$ for each i. What is the transmission cost of $(t_1, \ldots, t_n)$ to Bob?

  Solution: It is $nH(t_1|s_1) = n \log_2 4 = 2n$. Once s_i is given, then there are only 4 choices left for t_i and Alice only needs to send $\log_2 4$ bits to uniquely determine this value.

References

[CT06] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, second ed., Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006. MR 2239987 (2007h:00002)

[Sha48] C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J. **27** (1948), 379–423, 623–656. MR 0026286 (10,133e)

School of Computer Science, McGill University, Montréal, Canada
E-mail address: hatami@cs.mcgill.ca